

サポートベクトル 分類(8章)

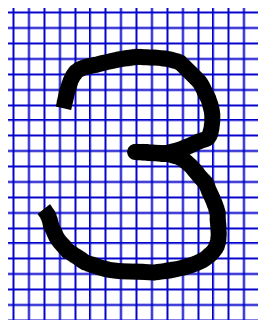
杉山将・本多淳也

sugi@k.u-tokyo.ac.jp, jhonda@k.u-tokyo.ac.jp

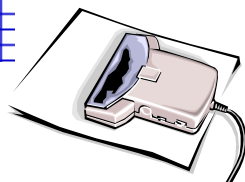
<http://www.ms.k.u-tokyo.ac.jp>

- 入力パターンをカテゴリに割り当てる
識別関数を構成する問題
 - 問題に合わせて人間が識別関数を設計
 - データから自動的に識別関数を学習

パターン



$x \in \mathbf{R}^d$



識別関数

$$y = f(x)$$



カテゴリ

3

$y \in \{1, 2, \dots, c\}$

統計的パターン認識

3

- **訓練標本**: 属するカテゴリが既知のパターン

$$\{(x_i, y_i)\}_{i=1}^n$$

$$x_i \in \mathbf{R}^d$$

$$y_i \in \{1, 2, \dots, c\}$$

- **統計的パターン認識**: 訓練標本の統計的な性質を利用して識別関数を学習する

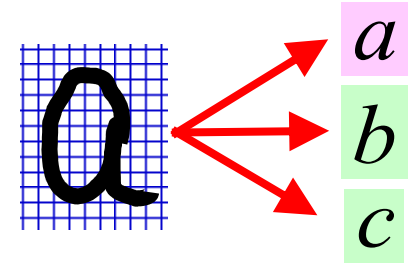
- **仮定**:

$$(x_i, y_i) \stackrel{\text{i.i.d.}}{\sim} p(x, y)$$

i.i.d. (independent and identically distributed)
独立に同一の分布に従う

理想的なパターン分類法

- **事後確率** $p(y|x)$: 与えられたパターン x がクラス y に属する確率



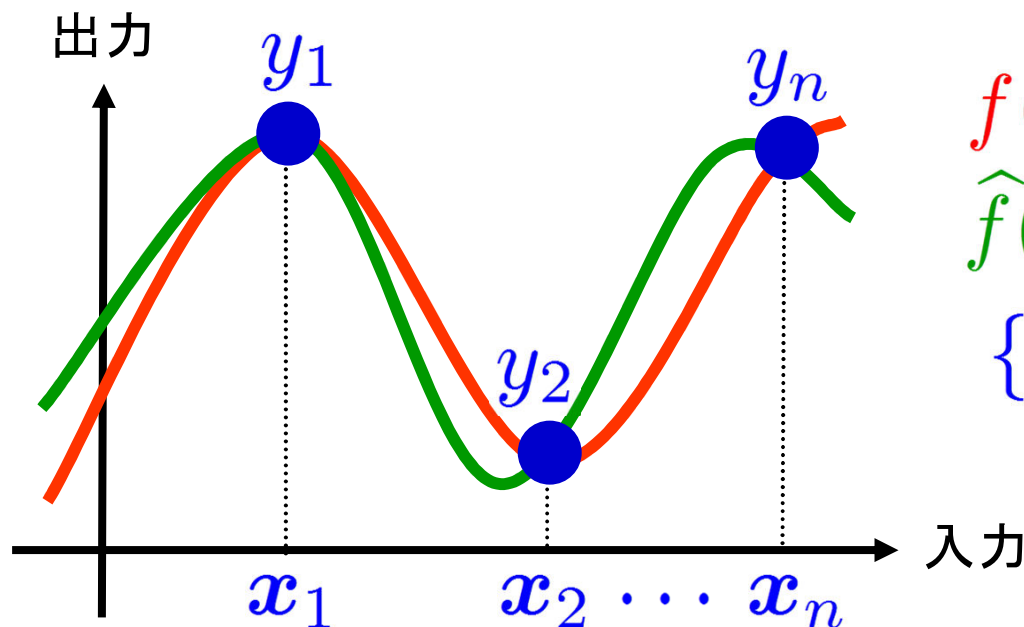
- 事後確率を最大にするカテゴリにパターンを分類すれば, パターンの誤識別率が最小になる.

$$f(x) = \arg \max_y p(y|x)$$

- 実際には事後確率は未知なので, 訓練標本から推定しなければならない.

識別モデルの学習 = 関数近似

5



$f(x)$: 学習したい真の関数

$\hat{f}(x)$: 学習結果の関数

$\{(x_i, y_i)\}_{i=1}^n$: 訓練標本

$y_i = f(x_i) (+\text{noise})$

訓練標本から真の関数にできるだけ近い関数を求める

パターン認識では $y \in \{1, 2, \dots, c\}$ であるが、
上記の図は $y \in \mathbb{R}$ (回帰) に対応している。

パラメータに関する線形モデル 6

■ 線形モデル:
$$f_{\theta}(\mathbf{x}) = \sum_{j=1}^b \theta_j \phi_j(\mathbf{x})$$

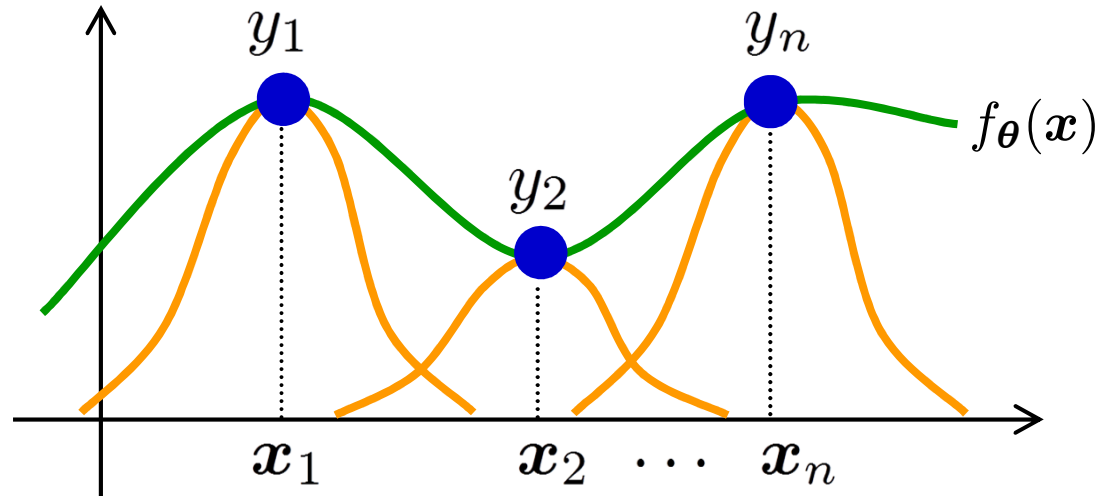
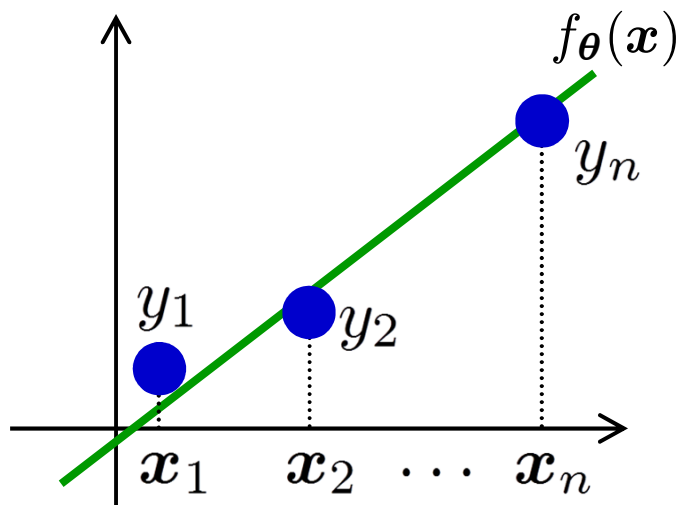
$\{\phi_j(\mathbf{x})\}_{j=1}^b$
: 基底関数

■ カーネルモデル:

ガウスカーネル

$$f_{\theta}(\mathbf{x}) = \sum_{j=1}^n \theta_j K(\mathbf{x}, \mathbf{x}_j)$$

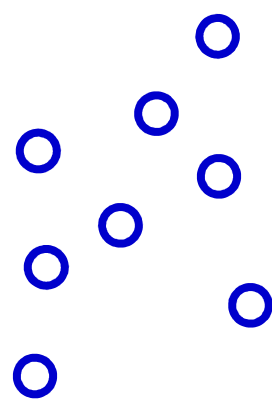
$$K(\mathbf{x}, \mathbf{c}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{c}\|^2}{2h^2}\right)$$



2クラス分類問題

- ラベル付き訓練データ: $\{(x_i, y_i)\}_{i=1}^n$
 - 入力 x は d 次元の実ベクトル $x \in \mathbb{R}^d$
 - 出力 y は2値のクラスラベル $y \in \{+1, -1\}$

クラス +1



分離境界

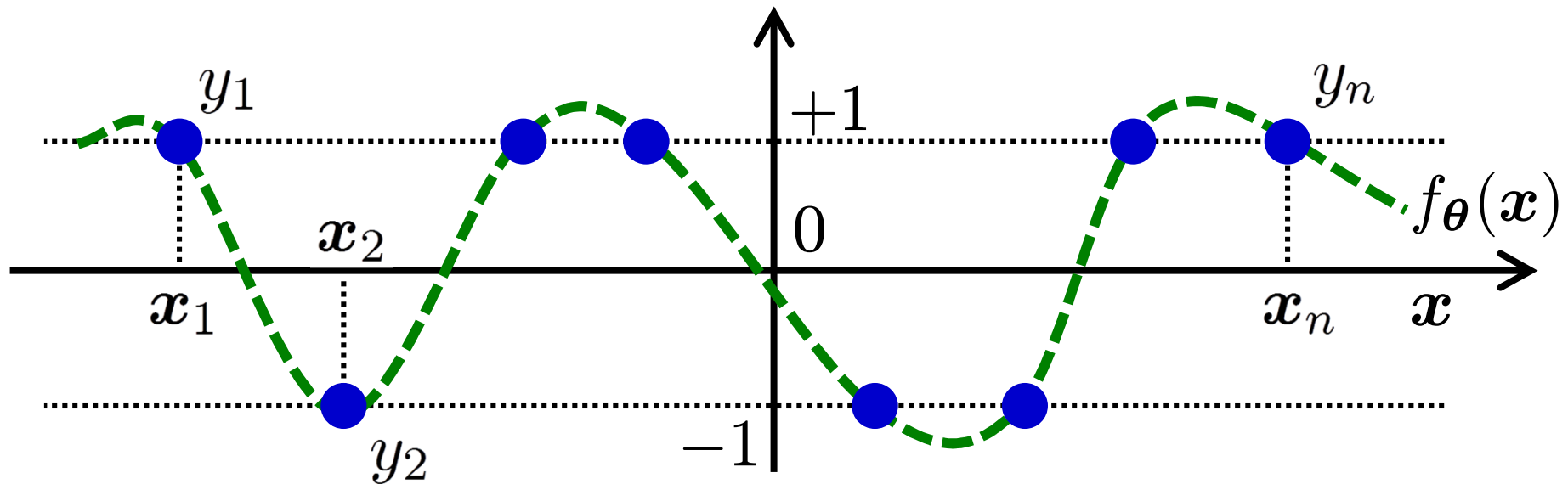
クラス-1

- クラス間の分離境界を求めたい

2クラスの分類問題

8

■ 2クラス分類問題は2値関数の近似問題と等価：



■ 回帰学習法が分類にも使える！

回帰学習による分類

■ パラメータを正則化最小二乗回帰で学習

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left[\frac{1}{2} \sum_{i=1}^n \left(f_{\boldsymbol{\theta}}(\mathbf{x}_i) - y_i \right)^2 + \frac{\lambda}{2} \|\boldsymbol{\theta}\|^2 \right]$$

$\lambda (\geq 0)$: 正則化パラメータ

■ テストパターンの分類:

$$\hat{y} = \operatorname{sign} \left(f_{\hat{\boldsymbol{\theta}}}(\mathbf{x}) \right) = \begin{cases} +1 & (f_{\hat{\boldsymbol{\theta}}}(\mathbf{x}) > 0) \\ 0 & (f_{\hat{\boldsymbol{\theta}}}(\mathbf{x}) = 0) \\ -1 & (f_{\hat{\boldsymbol{\theta}}}(\mathbf{x}) < 0) \end{cases}$$

0/1-損失関数とマージン

10

- 分類問題では, 学習した関数の符号だけが必要

$$\hat{y} = \text{sign} (f_{\hat{\theta}}(x))$$

- ℓ_2 -損失でなく, **0/1-損失**の方が自然

$$J_{0/1}(\theta) = \frac{1}{2} \sum_{i=1}^n (1 - \text{sign}(m_i))$$

$$m_i = f_{\theta}(x_i)y_i$$

マージン

- $\text{sign}(f_{\theta}(x_i)) = \text{sign}(y_i) \Rightarrow \text{sign}(m_i) = 1$
- $\text{sign}(f_{\theta}(x_i)) \neq \text{sign}(y_i) \Rightarrow \text{sign}(m_i) = -1$

- $J_{0/1}(\theta)$ は**誤分類標本数**に相当.

0/1-損失関数とマージン

11

$$J_{0/1}(\theta) = \frac{1}{2} \sum_{i=1}^n (1 - \text{sign}(m_i))$$

■ 0/1-損失は誤分類標本数に対応

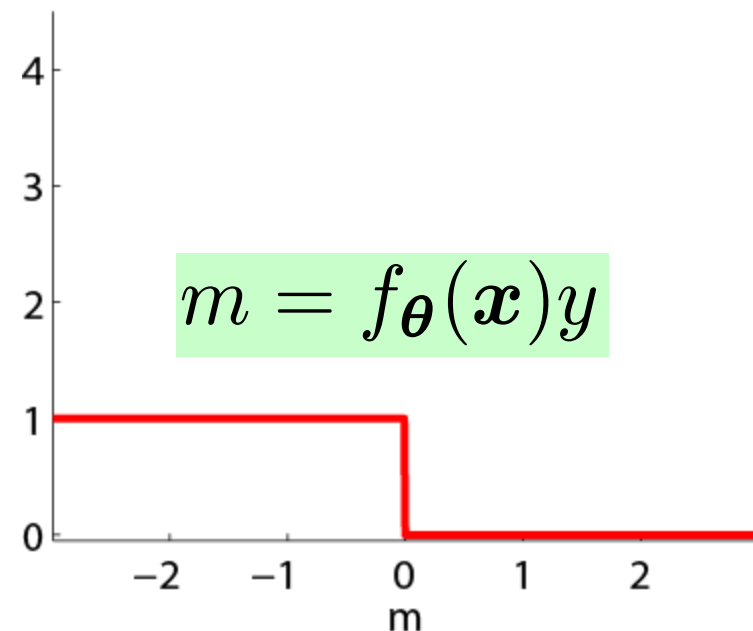
- マージンが正なら誤差0
- マージンが負なら誤差1

■ 分類の損失としては理想的

- しかし傾きを持たない
離散的な関数

■ 0/1-損失の最小化はNP困難

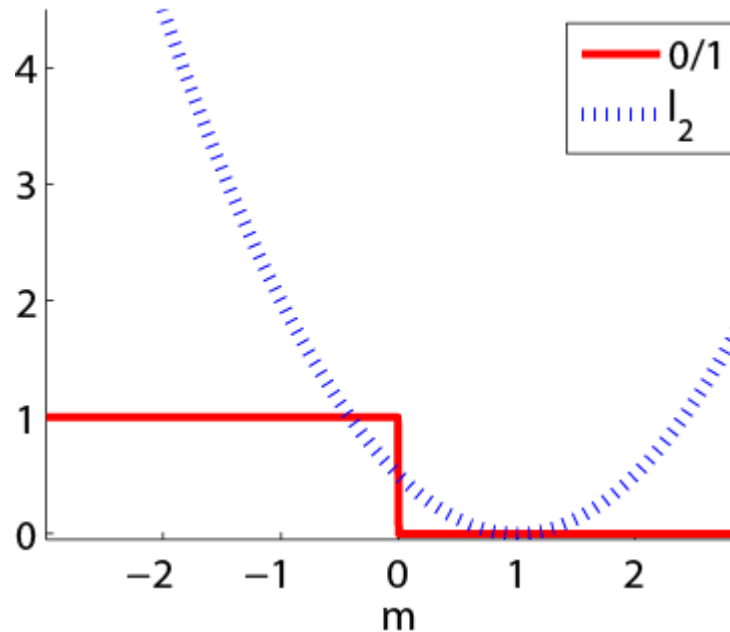
- 現実的な時間では不可能



ℓ_2 -損失とマージン

$$\frac{1}{2} \left(f_{\theta}(x) - y \right)^2 = \frac{1}{2} (1 - m)^2$$

$$m = f_{\theta}(x)y$$

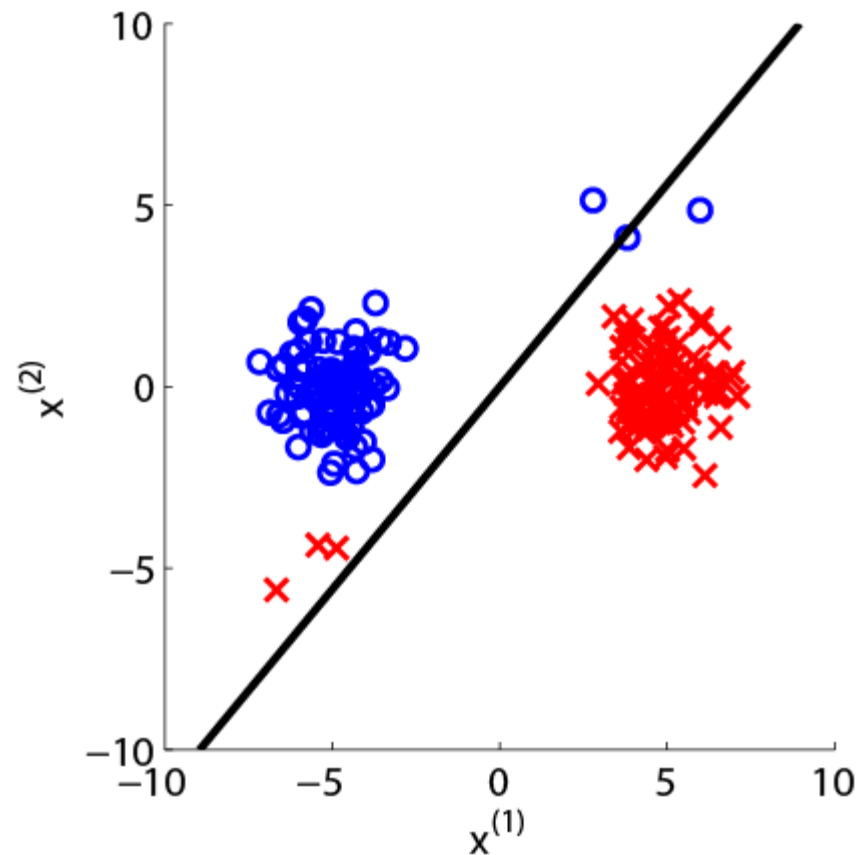


- ℓ_2 -損失関数は連続関数で扱いやすい
- 負のマージンを正(+1)にしようとする
- 正の大きいマージンを+1に減らそうとする

ℓ_2 -損失の問題点

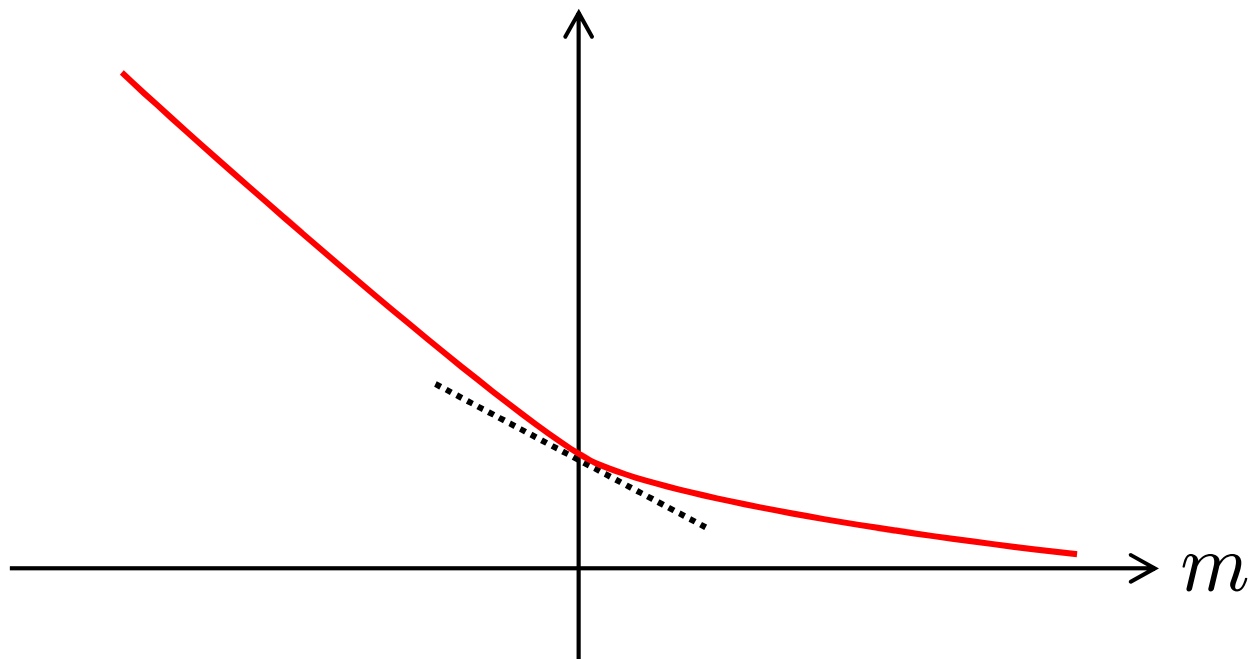
13

- 正の大きいマージンを+1に減らそうとすることにより, 下記のデータを正しく分離できない



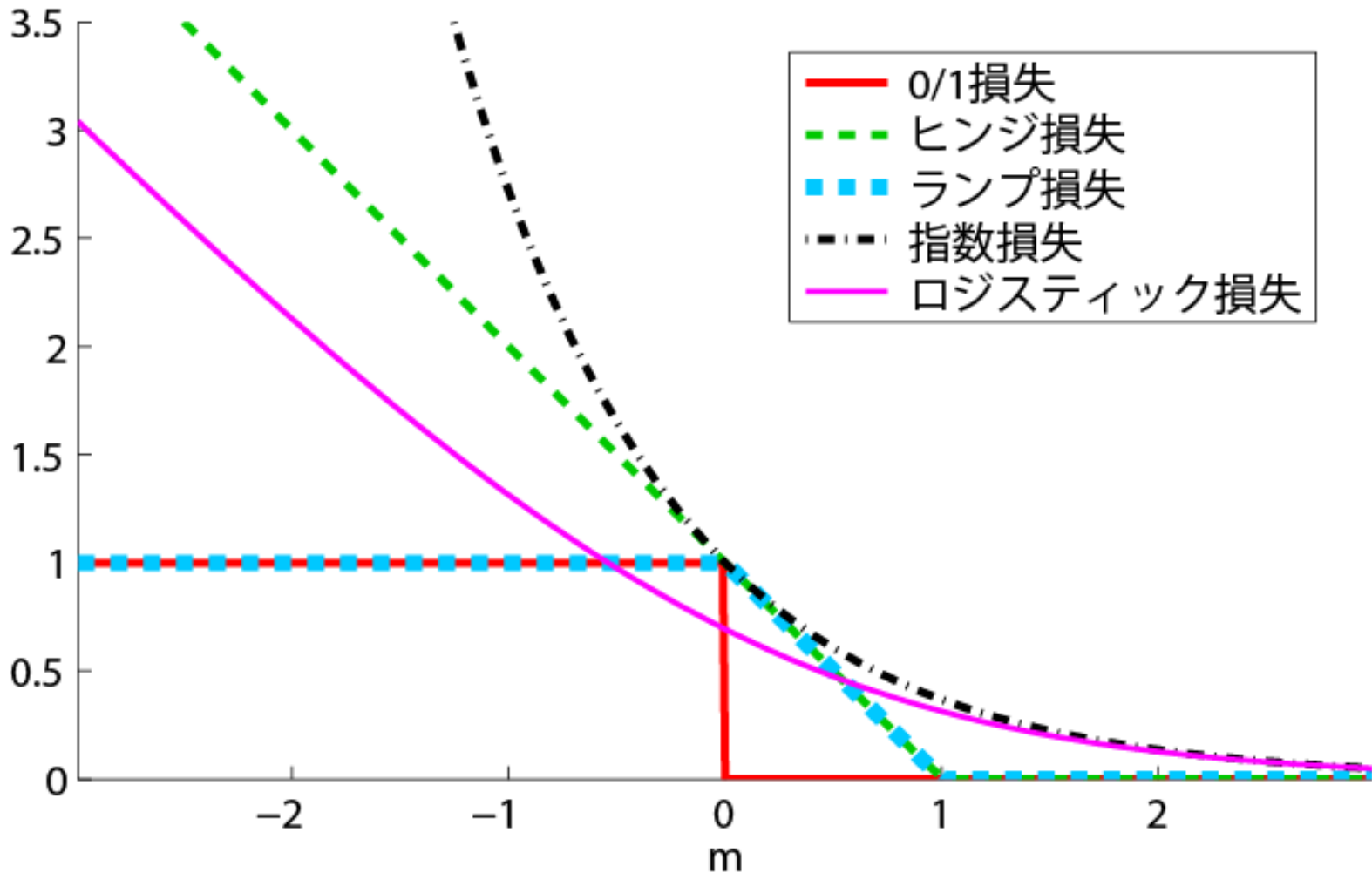
■ 0/1-損失の代理として使う損失は、
単調非増加で $m = 0$ での傾きが
負のものがよい

- 負のマージンを正にしようとする
- 正のマージンは減らさない



代理損失

■ 機械学習では、様々な代理損失が用いられる



講義の流れ

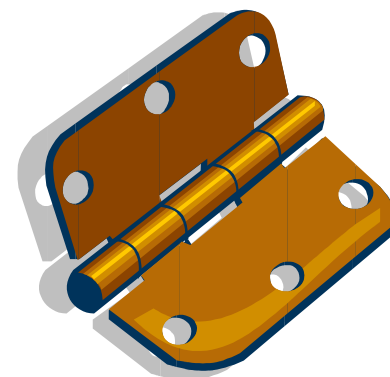
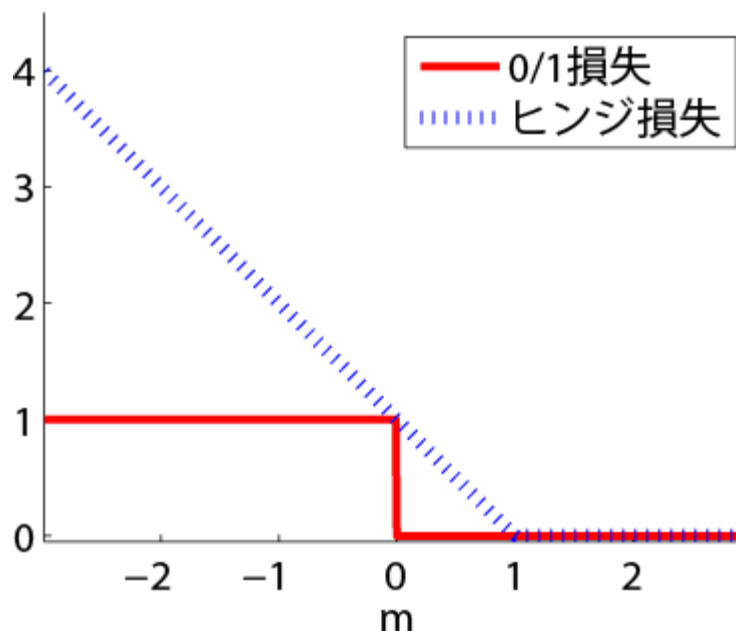


1. ヒンジ損失(8章)
2. サポートベクトル分類の元々の導出(8章)
3. サポートベクトル分類の解の性質(8章)

- 0/1-損失の代理としてヒンジ損失を用いる

$$\sum_{i=1}^n \max(0, 1 - m_i)$$

$$m_i = f_{\theta}(x_i)y_i$$



サポートベクトル分類

- ヒンジ損失と ℓ_2 -制約の和を最小にする

$$\min_{\theta \in \mathbb{R}^b} \left[\sum_{i=1}^n \max \left(0, 1 - f_{\theta}(\mathbf{x}_i) y_i \right) + \frac{\lambda}{2} \theta^{\top} R \theta \right]$$

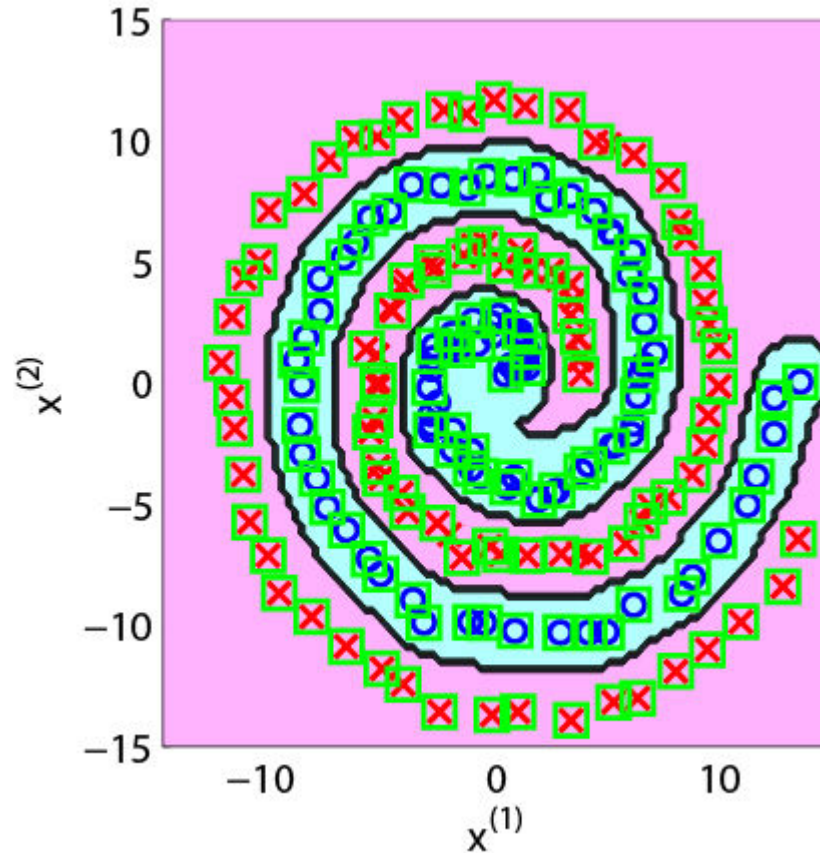
- 特に以下の場合をサポートベクトル分類器とよぶ

- $$f_{\theta}(\mathbf{x}) = \sum_{j=1}^n \theta_j K(\mathbf{x}, \mathbf{x}_j) + \theta_0$$

- $$R = K \quad K_{i,j} = K(\mathbf{x}_i, \mathbf{x}_j)$$

- 優れたソフトウェアが多数公開されている:

<http://www.support-vector-machines.org/>



- 複雑なデータもきちんと分離できている

劣勾配法

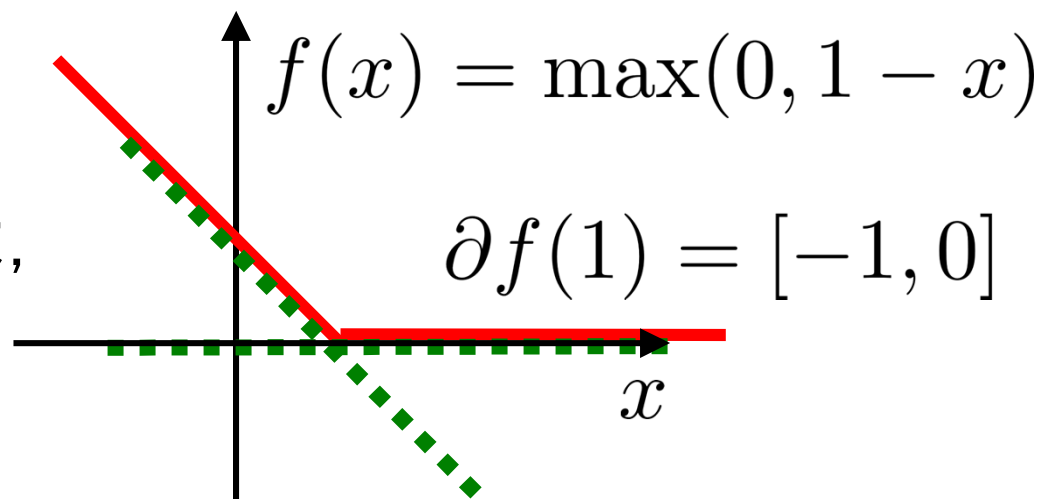
- 凸関数 f の x' での劣勾配(sub-gradient)とは、全ての $x \in \mathbb{R}^d$ に対して次式を満たす ξ :

$$f(x) \geq f(x') + \xi^\top (x - x')$$

- f が微分可能なとき, $\xi = \nabla f(x')$
- 上式を満たす ξ 全体を $\partial f(x')$ で表し
劣微分(sub-differential)とよぶ

■ 劣勾配法:

- 勾配法において、
微分不可能な点では、
劣微分のどれかの
値を用いる



$$\min_{\boldsymbol{\theta} \in \mathbb{R}^b} \left[\sum_{i=1}^n \max \left(0, 1 - f_{\boldsymbol{\theta}}(\mathbf{x}_i) y_i \right) + \frac{\lambda}{2} \boldsymbol{\theta}^\top \mathbf{R} \boldsymbol{\theta} \right]$$

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = \sum_{j=1}^n \theta_j \exp \left(-\frac{\|\mathbf{x} - \mathbf{x}_j\|^2}{2h^2} \right)$$

■ ヒンジ損失の $\boldsymbol{\theta}$ に関する劣微分

$$\partial \max \left(0, 1 - f_{\boldsymbol{\theta}}(\mathbf{x}_i) y_i \right)$$

を求めよ

$$\partial \max (0, 1 - z) = \begin{cases} -1 & z < 1 \\ [-1, 0] & z = 1 \\ 0 & z > 1 \end{cases}$$

$$t_j = \partial_j \max \left(0, 1 - f_{\theta}(\mathbf{x}_i)y_i \right) \quad f_{\theta}(\mathbf{x}_i) = \sum_{j=1}^n \theta_j \exp \left(-\frac{\|\mathbf{x} - \mathbf{x}_j\|^2}{2h^2} \right)$$

■ $1 - f_{\theta}(\mathbf{x}_i)y_i > 0$ のとき $t_j = -y_i \exp \left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2h^2} \right)$

■ $1 - f_{\theta}(\mathbf{x}_i)y_i = 0$ かつ $y_i = +1$ のとき

$$t_j = \left[-\exp \left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2h^2} \right), 0 \right]$$

■ $1 - f_{\theta}(\mathbf{x}_i)y_i = 0$ かつ $y_i = -1$ のとき

$$t_j = \left[0, \exp \left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2h^2} \right) \right]$$

■ $1 - f_{\theta}(\mathbf{x}_i)y_i < 0$ のとき $t_j = 0$

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \varepsilon \left(\sum_{i=1}^n \partial \max \left(0, 1 - f_{\boldsymbol{\theta}}(\mathbf{x}_i)y_i \right) + \lambda \mathbf{R}\boldsymbol{\theta} \right)$$

$\varepsilon > 0$: ステップ幅

■ 劣勾配:

$$1 - f_{\boldsymbol{\theta}}(\mathbf{x}_i)y_i > 0 \quad \longrightarrow \quad \partial_j \max \left(0, 1 - f_{\boldsymbol{\theta}}(\mathbf{x}_i)y_i \right) \\ = -y_i \exp \left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2h^2} \right)$$

$$1 - f_{\boldsymbol{\theta}}(\mathbf{x}_i)y_i \leq 0 \quad \longrightarrow \quad \partial_j \max \left(0, 1 - f_{\boldsymbol{\theta}}(\mathbf{x}_i)y_i \right) = 0$$

講義の流れ



1. ヒンジ損失(8章)
2. サポートベクトル分類の元々の導出(8章)
3. サポートベクトル分類の解の性質(8章)

サポートベクトル分類器の 正統派の導出

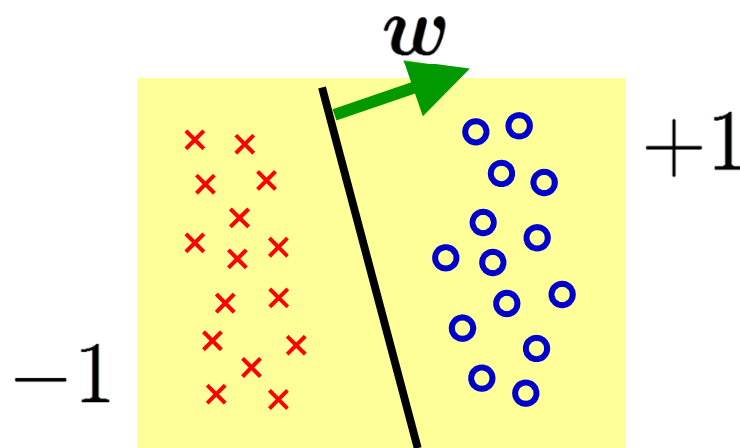
- 超平面分類器
- VC理論
- マージン最大化
- ソフトマージン
- カーネルトリック

- 標本空間を超平面で分離する.

$$f_{\boldsymbol{w},b}(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$$

$$\boldsymbol{w} = (w^{(1)}, \dots, w^{(d)})^\top$$

$$\boldsymbol{x} = (x^{(1)}, \dots, x^{(d)})^\top$$




find \boldsymbol{w}, b

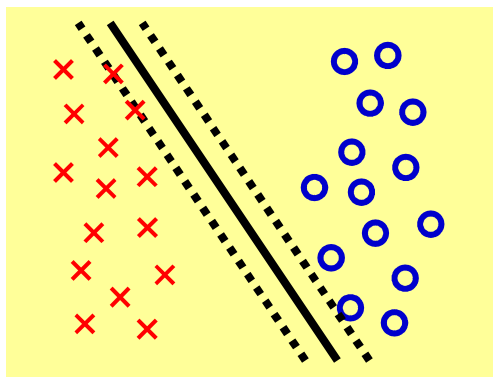
such that $y_i f_{\boldsymbol{w},b}(\boldsymbol{x}_i) \geq 1$ for $i = 1, \dots, n$.

マージン

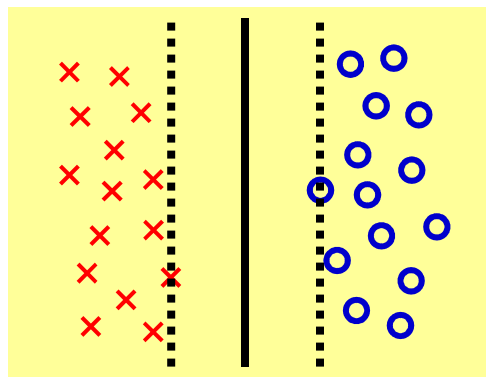
■ マージン: 二つのクラスの“隙間”の大きさ

マージン
 $1/\|w\|$

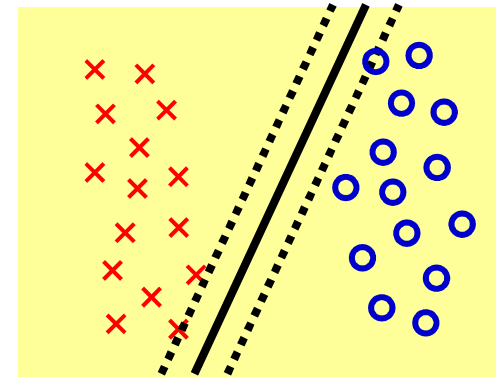




マージン: 小



マージン: 大



マージン: 小

$$f_{w,b}(x) = w^\top x + b$$

$$y_i f_{w,b}(x_i) \geq 1 \quad \text{for } i = 1, \dots, n.$$

■ 汎化誤差: $R[\hat{f}] = \int \int I(\hat{f}(\mathbf{x}) \neq y) p(\mathbf{x}, y) d\mathbf{x} dy$

■ 経験誤差: $R_{\text{emp}}[\hat{f}] = \frac{1}{n} \sum_{i=1}^n I(\hat{f}(\mathbf{x}_i) \neq y_i)$

$$I(a \neq b) = \begin{cases} 0 & (a = b) \\ 1 & (a \neq b) \end{cases}$$

■ 汎化誤差の確率的上界 (“**VCバウンド**”)

$$R[\hat{f}] \leq R_{\text{emp}}[\hat{f}] + \sqrt{\frac{1}{n} \left(V \left(\log \frac{2n}{V} + 1 \right) + \log \frac{4}{\delta} \right)}$$

V : VC次元 (分類器の複雑さ)

with probability $1 - \delta$

Vapnik-Chevonenkis理論 (続き) 29

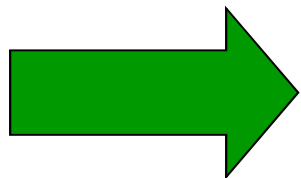
■ VCバウンド:

$$R[\hat{f}] \leq R_{\text{emp}}[\hat{f}] + \underbrace{\sqrt{\frac{1}{n} \left(V \left(\log \frac{2n}{V} + 1 \right) + \log \frac{4}{\delta} \right)}}_{\text{VC次元 } V \text{ (} V < n \text{) の減少に対して単調減少}}$$

■ 標本が線形分離可能なとき, 経験誤差はゼロ:

$$R_{\text{emp}}[\hat{f}] = 0$$

■ マージンが大きいほど, VC次元は小さい.

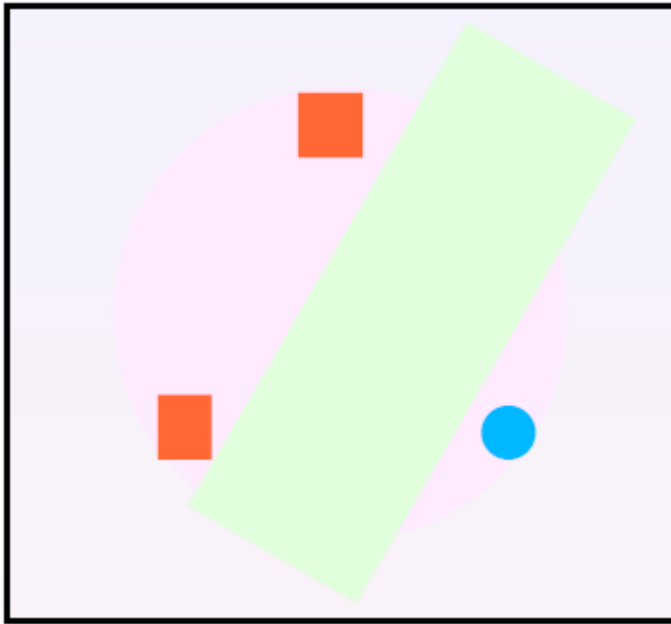


マージンが最大の超平面識別器が,
VC理論においては最適

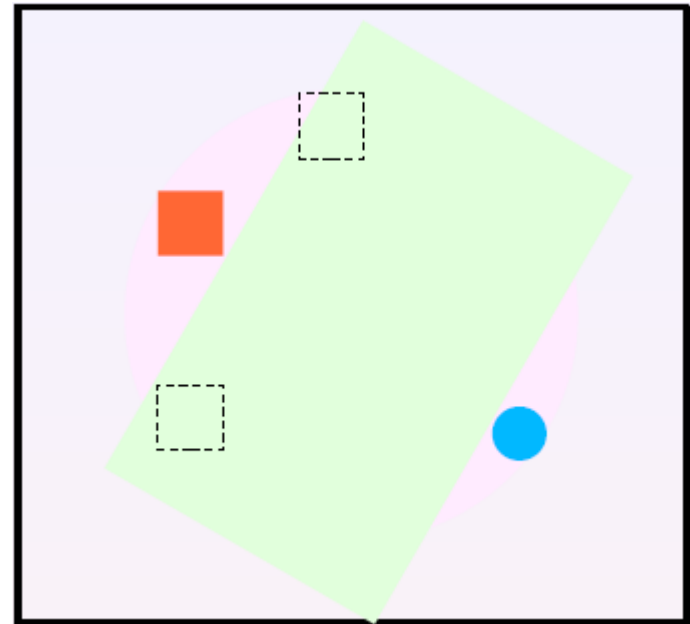
マージンとVC次元

- 一定ノルム内の点の数が多い場合, 大きなマージンをもつ超平面識別器が存在できない

マージン小



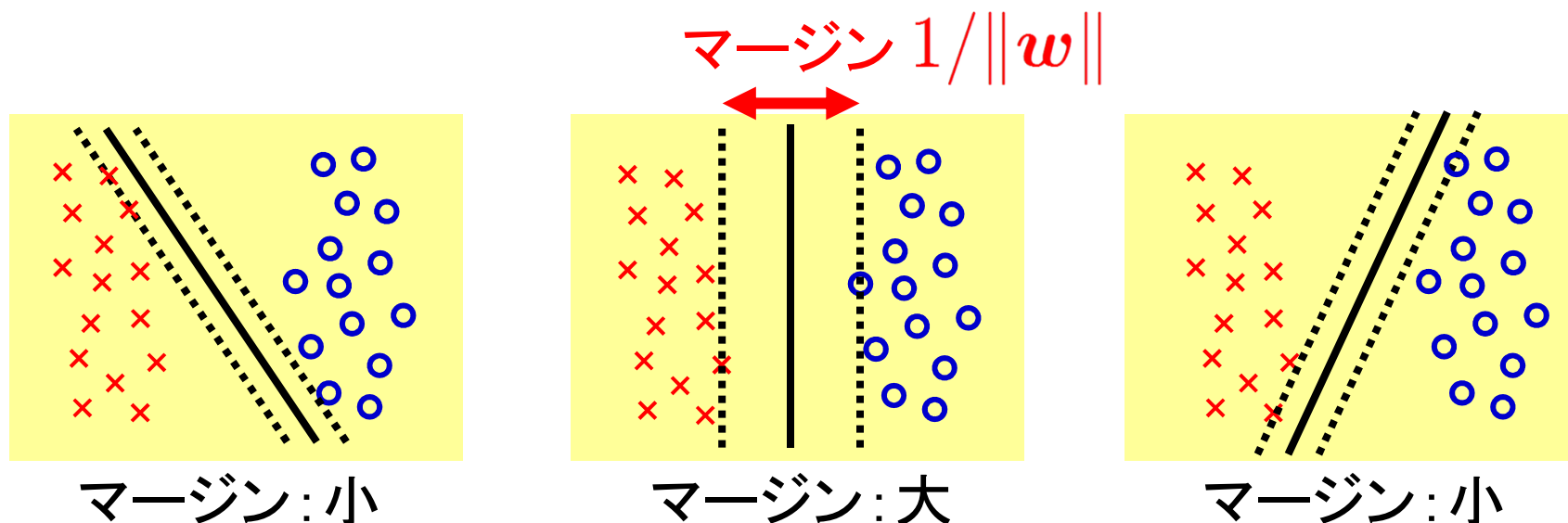
マージン大



[<http://www.ide.titech.ac.jp/~yamasita/yylab/06nhk.pdf>]

最適超平面分類器

- マージンが最大になるように二つのクラスを超平面で分ける.



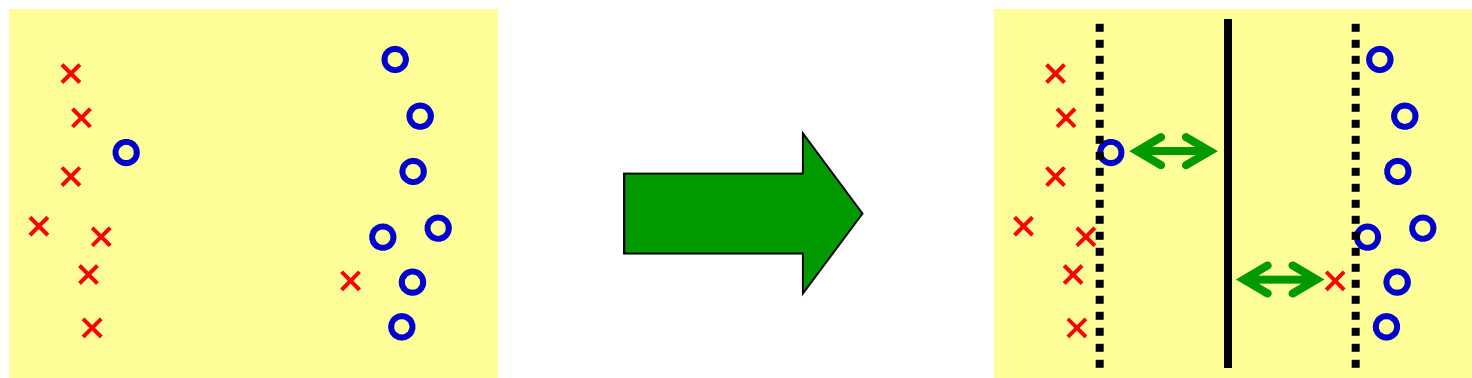
$$\min_{w,b} \|w\|^2$$

$$f_{w,b}(x) = w^\top x + b$$

$$\text{subject to } y_i f_{w,b}(x_i) \geq 1 \quad \text{for } i = 1, \dots, n.$$

ソフトマージン

- 標本が線形分離可能でないときはマージンが定義できない.
- 少しの**誤差** ξ_i を許す.



$$\min_{\mathbf{w}, b, \boldsymbol{\xi}} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \xi_i$$

$$f_{\mathbf{w}, b}(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$$

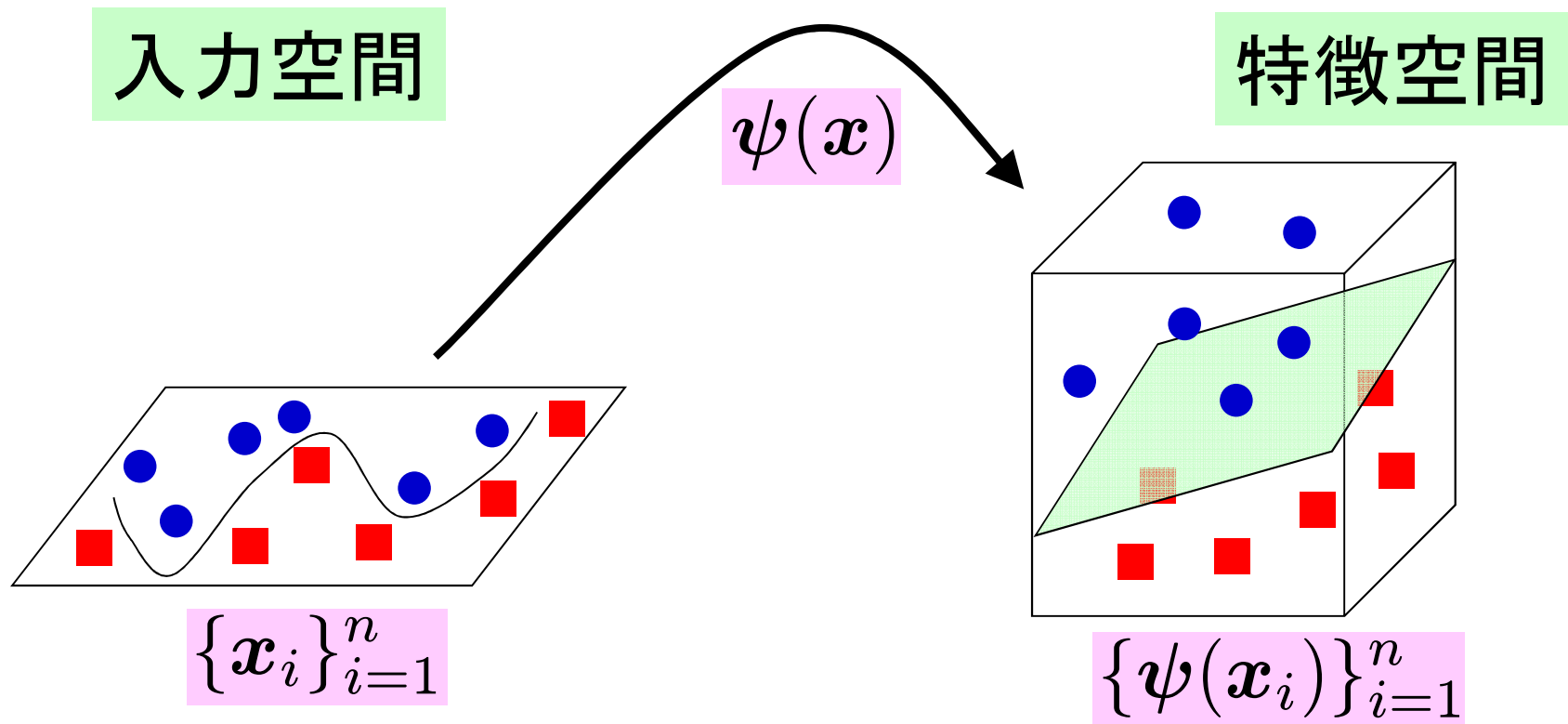
subject to $y_i f_{\mathbf{w}, b}(\mathbf{x}_i) \geq 1 - \xi_i$

$$\xi_i \geq 0 \quad \text{for } i = 1, \dots, n.$$

非線形への拡張

33

- 非線形関数 $\psi(x)$ で標本を特徴空間へ写像し、特徴空間内でマージン最大の超平面を求める。



カーネルトリック

- 特徴空間内での内積をカーネル関数で計算:

$$\psi(x_i)^\top \psi(x_j) = K(x_i, x_j)$$

$$\forall x, x', \quad K(x, x') \geq 0$$

例えばガウシアンカーネル

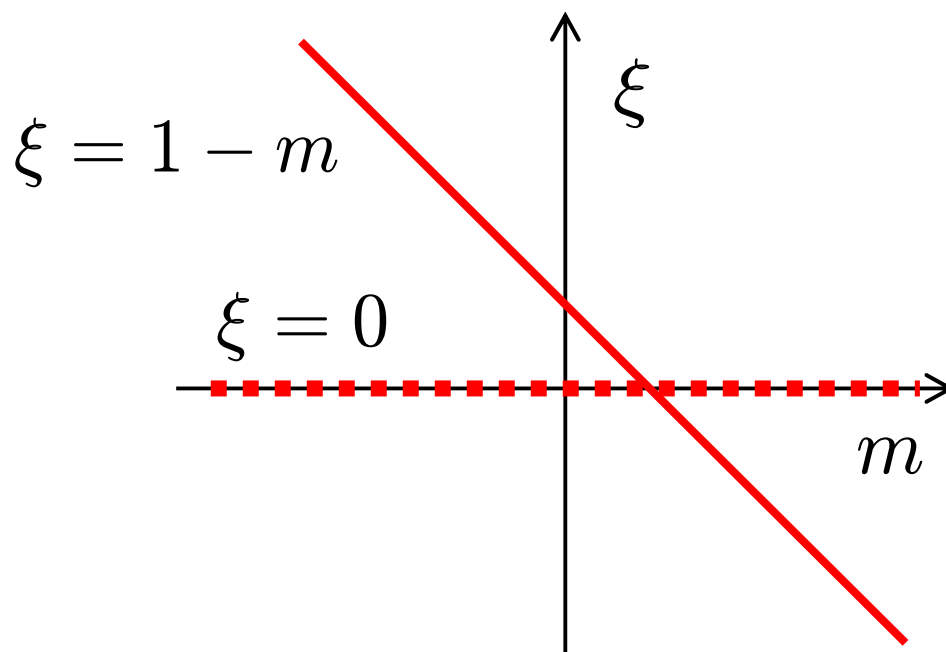
$$K(x, x') = \exp \left(-\|x - x'\|^2 / (2h^2) \right)$$

- 内積のみで表される線形アルゴリズムは、そのまま非線形に拡張できる.
 - 例: サポートベクトルマシン, 主成分分析, 線形判別分析, K平均クラスタリングなど

ヒンジ損失との関係

- ヒンジ損失は以下のように変形できる

$$\max\{0, 1 - m\} = \min_{\xi} \xi \quad \text{subject to } \xi \geq 1 - m, \xi \geq 0$$



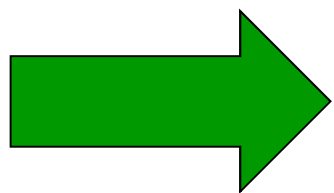
$$\max\{0, 1 - m\} = \min_{\xi} \xi \quad \text{subject to } \xi \geq 1 - m, \xi \geq 0$$

■ ξ_i を消去すると制約なし最適化問題に変形できる:

$$\min_{\mathbf{w}, b, \xi} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \xi_i$$

$$\text{subject to } y_i f_{\mathbf{w}, b}(\mathbf{x}_i) \geq 1 - \xi_i$$

$$\xi_i \geq 0 \quad \text{for } i = 1, \dots, n$$



$$\min_{\mathbf{w}, b} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \max\left(0, 1 - y_i f_{\mathbf{w}, b}(\mathbf{x}_i)\right)$$

ヒンジ損失との関係

$$\min_{\mathbf{w}, b} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \max \left(0, 1 - y_i f_{\mathbf{w}, b}(\mathbf{x}_i) \right)$$

$$f_{\mathbf{w}, b}(\mathbf{x}) = \mathbf{w}^\top \boldsymbol{\psi}(\mathbf{x}) + b$$

■ $\mathbf{w} = \sum_{j=1}^n \theta_j \boldsymbol{\psi}(\mathbf{x}_j)$ とおけばヒンジ損失最小化と等価

$$\min_{\boldsymbol{\theta}} \left[\sum_{i=1}^n \max \left(0, 1 - f_{\boldsymbol{\theta}}(\mathbf{x}_i) y_i \right) + \lambda \boldsymbol{\theta}^\top \mathbf{K} \boldsymbol{\theta} \right]$$

$$C = 1/\lambda$$

$$\boldsymbol{\psi}(\mathbf{x}_i)^\top \boldsymbol{\psi}(\mathbf{x}_j) = K(\mathbf{x}_i, \mathbf{x}_j)$$

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = \sum_{j=1}^n \theta_j K(\mathbf{x}, \mathbf{x}_j) + \gamma$$

講義の流れ



1. ヒンジ損失(8章)
2. サポートベクトル分類の元々の導出(8章)
3. サポートベクトル分類の解の性質(8章)

■ 主問題(primal problem):

$$\min_{x \in \mathcal{X}} f(x) \quad \text{subject to } g(x) = 0, \quad h(x) \leq 0$$

$$g(x) = (g_1(x), \dots, g_m(x))^{\top}$$

$$h(x) = (h_1(x), \dots, h_n(x))^{\top}$$

■ 双対問題(dual problem):

- 主問題の最適値の下界を最大化する問題
- 制約条件から目的関数を作る
- 目的関数から制約条件を作る

$$f^* = \min_{x \in \mathcal{X}} f(x) \quad \text{subject to } g(x) = 0, \quad h(x) \leq 0$$

■ ラグランジュ関数(Lagrangian):

$$L(x, \lambda, \mu) = f(x) + \lambda^\top g(x) + \mu^\top h(x)$$

■ ラグランジュ双対問題(Lagrange dual problem):

$$f^* = \max_{\lambda, \mu} \inf_{x \in \mathcal{X}} L(x, \lambda, \mu) \\ \text{subject to } \mu \geq 0$$

- 凸最適化問題に対しては最適解が一致
- 双対問題の方が制約が単純

$$\min_{\mathbf{w}, \xi} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \xi_i$$

$$f_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$$

(単純化のため
+bを省略)

$$\text{subject to } y_i f_{\mathbf{w}}(\mathbf{x}_i) \geq 1 - \xi_i$$

$$\xi_i \geq 0 \quad \text{for } i = 1, \dots, n$$

■ サポートベクトルマシンのラグランジュ関数
 $L(\mathbf{w}, \xi, \alpha, \beta)$ を求めよ

● $\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x})$ subject to $g(\mathbf{x}) = 0, h(\mathbf{x}) \leq 0$

のラグランジュ関数は

$$L(\mathbf{x}, \lambda, \mu) = f(\mathbf{x}) + \lambda^\top g(\mathbf{x}) + \mu^\top h(\mathbf{x})$$

$$L(\boldsymbol{w}, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2} \|\boldsymbol{w}\|^2 + C \sum_{i=1}^n \xi_i$$

$$- \sum_{i=1}^n \alpha_i (y_i \boldsymbol{w}^\top \boldsymbol{x}_i - 1 + \xi_i) - \sum_{i=1}^n \beta_i \xi_i$$

■ サポートベクトルマシンの双対最適化問題

$$\max_{\alpha, \beta} \inf_{w, \xi} L(w, \xi, \alpha, \beta)$$

subject to $\alpha \geq 0$ and $\beta \geq 0$

から w, ξ, β を消去せよ

■ ヒント: $\frac{\partial L}{\partial w} = 0, \frac{\partial L}{\partial \xi} = 0$

$$L(w, \xi, \alpha, \beta) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i$$

$$- \sum_{i=1}^n \alpha_i (y_i w^\top x_i - 1 + \xi_i) - \sum_{i=1}^n \beta_i \xi_i$$

$$\blacksquare \frac{\partial L}{\partial \mathbf{w}} = 0 \implies \mathbf{w} = \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i$$

$$\blacksquare \frac{\partial L}{\partial \xi_i} = 0 \implies \alpha_i + \beta_i = C, \forall i = 1, \dots, n$$

■ これらより

$$\hat{\boldsymbol{\alpha}} = \operatorname{argmax}_{\boldsymbol{\alpha}} \left[\sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j \mathbf{x}_i^{\top} \mathbf{x}_j \right]$$

subject to $0 \leq \alpha_i \leq C$ for $i = 1, \dots, n$

- 双対問題は二次計画(quadratic program; QP):

$$\hat{\alpha} = \operatorname{argmax}_{\alpha} \left[\sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j \mathbf{x}_i^{\top} \mathbf{x}_j \right]$$

subject to $0 \leq \alpha_i \leq C$ for $i = 1, \dots, n$

- 主問題の解:

$$\hat{\mathbf{w}} = \sum_{i=1}^n \hat{\alpha}_i y_i \mathbf{x}_i$$

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \implies \mathbf{w} = \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i$$

最適性条件

$$\min_{x \in \mathcal{X}} f(x) \quad \text{subject to } g(x) = 0, \quad h(x) \leq 0$$

$$g(x) = (g_1(x), \dots, g_m(x))^{\top}$$

$$h(x) = (h_1(x), \dots, h_n(x))^{\top}$$

■ 最適解の必要条件 (凸の場合は必要十分条件):

$$\bullet \quad \nabla f(x^*) + \lambda^{*\top} \nabla g(x^*) + \mu^{*\top} \nabla h(x^*) = 0$$

$$\bullet \quad g(x^*) = 0$$

$$\bullet \quad h(x^*) \leq 0$$

$$\bullet \quad \mu^* \geq 0$$

$$\bullet \quad \mu_i^* h_i(x^*) = 0, \quad i = 1, \dots, n \quad \leftarrow \text{相補性条件}$$

カルーシュ・キューン・タッカー
(KKT)条件

(Karush-Kuhn-Tucker conditions)

$$\min_{\mathbf{w}, \xi} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \xi_i \quad f_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$$

$$\text{subject to } y_i f_{\mathbf{w}}(\mathbf{x}_i) \geq 1 - \xi_i$$

$$\xi_i \geq 0 \quad \text{for } i = 1, \dots, n$$

- サポートベクトルマシンの双対最適化問題の相補性条件は

$$\alpha_i (y_i \mathbf{w}^\top \mathbf{x}_i - 1 + \xi_i) = 0 \quad \beta_i \xi_i = 0$$

$$\text{for } i = 1, \dots, n$$

相補性条件(続き)

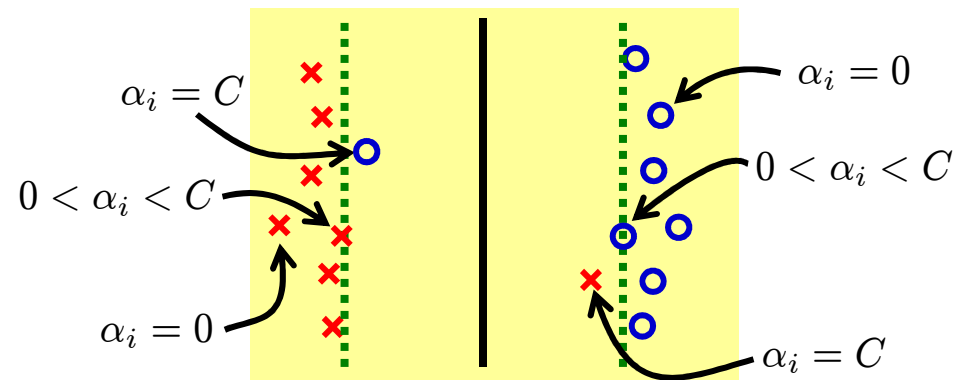
48

$$\alpha_i(y_i \mathbf{w}^\top \mathbf{x}_i - 1 + \xi_i) = 0 \quad \beta_i \xi_i = 0$$

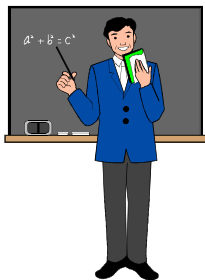
■ 相補性条件より, 以下の性質が成り立つ:

1. $\alpha_i = 0 \implies y_i \mathbf{w}^\top \mathbf{x}_i \geq 1$
2. $0 < \alpha_i < C \implies y_i \mathbf{w}^\top \mathbf{x}_i = 1$
3. $\alpha_i = C \implies y_i \mathbf{w}^\top \mathbf{x}_i \leq 1$
4. $y_i \mathbf{w}^\top \mathbf{x}_i > 1 \implies \alpha_i = 0$
5. $y_i \mathbf{w}^\top \mathbf{x}_i < 1 \implies \alpha_i = C$

導出は宿題

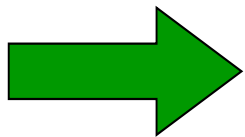


講義の流れ



1. ヒンジ損失(8章)
2. サポートベクトル分類の元々の導出(8章)
3. サポートベクトル分類の解の性質(8章)

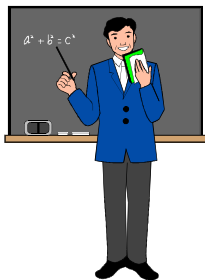
- 0/1-損失の代理としては**ヒンジ損失**が適切
- サポートベクトル分類は, 元々は**マージン最大化原理**に基づいた学習法として導出
- **分類問題**:
 - 入力に関する線形モデル + 最小二乗学習
= **フィッシャーの線形判別分析**
 - カーネルモデル + 正則化ヒンジ損失最小化
= **サポートベクトルマシン**



良い方法には様々な解釈がある

次回の予告

- 確率的分類(10章)
- 系列データの分類(11章)

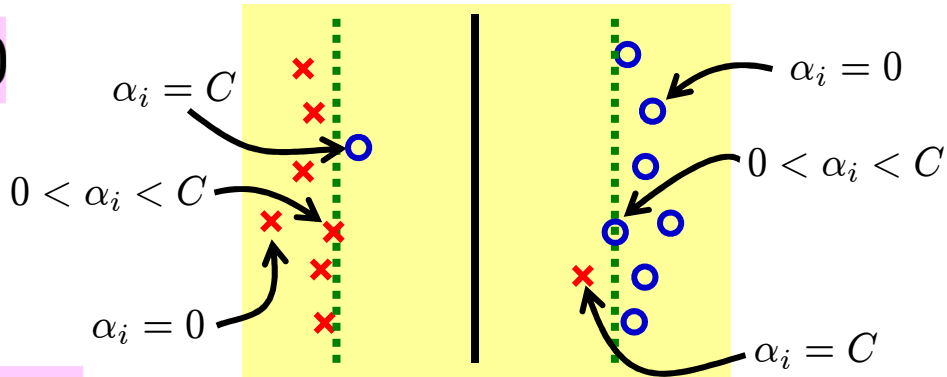


相補性条件

- $\alpha_i(y_i \mathbf{w}^\top \mathbf{x}_i - 1 + \xi_i) = 0$
- $\beta_i \xi_i = 0$

より, 以下の性質を示せ

1. $\alpha_i = 0 \implies y_i \mathbf{w}^\top \mathbf{x}_i \geq 1$
2. $0 < \alpha_i < C \implies y_i \mathbf{w}^\top \mathbf{x}_i = 1$
3. $\alpha_i = C \implies y_i \mathbf{w}^\top \mathbf{x}_i \leq 1$
4. $y_i \mathbf{w}^\top \mathbf{x}_i > 1 \implies \alpha_i = 0$
5. $y_i \mathbf{w}^\top \mathbf{x}_i < 1 \implies \alpha_i = C$



ヒント: 以下の条件を利用する

$$y_i \mathbf{w}^\top \mathbf{x}_i - 1 + \xi_i \geq 0 \quad \xi_i \geq 0 \quad \alpha_i + \beta_i = C$$

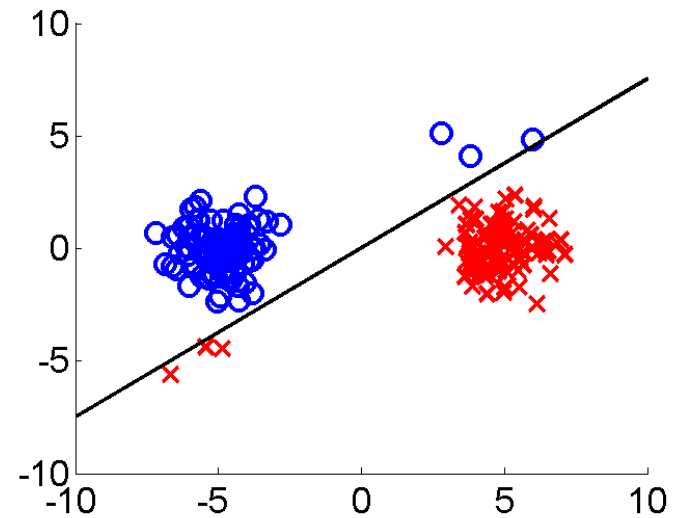
宿題2

53

■ 線形モデル

$$f_{w,b}(x) = w^T x + b$$

に対するサポートベクトルマシンの
劣勾配アルゴリズムを実装せよ



```
clear all; rand('state',0); randn('state',0);
n=200; x=[randn(1,n/2)-5 randn(1,n/2)+5; randn(1,n)]';
x(:,3)=1;
y=[ones(n/2,1);-ones(n/2,1)]; y(1:3)=-1; y(n/2+1:n/2+3,1)=1;
x(1:3,2)=x(1:3,2)-5; x(n/2+1:n/2+3,2)=x(n/2+1:n/2+3,2)+5;
```

<<< From x and y, learn 3-dimensional vector w >>>

```
figure(1); clf; hold on; axis([-10 10 -10 10]);
plot(x(y==1,1),x(y==1,2),'bo');
plot(x(y==-1,1),x(y==-1,2),'rx');
plot([-10 10],-(w(3)+[-10 10]*w(1))/w(2),'k-');
```